

Лабораторная работа №5

Метрики качества классификации

Цель работы: рассмотреть различные метрики качества классификации, входящих в состав библиотеки scikit-learn.

Теоретический материал

В предыдущих работах были рассмотрены различные алгоритмы классификации. Но при этом пока не оценивалось качество алгоритма, чтобы понять хорошо или плохо он работает на той или иной выборке.

Итак, пусть дана выборка X , для каждого вектора которой мы знаем бинарный ответ – ноль или один, нулевой или первый класс. И дан некоторый алгоритм «а», качество которого мы хотим измерить на этой выборке.

$X^l = \{x_1, \dots, x_l\}$ – выборка

$y_i = y(x_i) \in \{0,1\}, i = 1, \dots, l$ – известные бинарные ответы

$a: X \rightarrow Y$ – алгоритм, решающая функция, приближающая y на всём множестве объектов X

Первое, что можно сделать, это посчитать на скольких объектах мы даем правильный ответ, и поделить это на размер выборки. Это метрика называется доля правильных ответов (Accuracy). Эта метрика:

- Соответствует интуитивным представлениям о качестве классификации
- Имеет проблемы с интерпретацией на несбалансированных выборках

Каждый объект характеризуется двумя числами – тем, какой правильный ответ на нем, и тем, какой ответ дает алгоритм. В зависимости от сочетания этих двух чисел, можно разделить все объекты на четыре категории. Первая – это «верные срабатывания» (True Positive). К ним относятся те объекты, на которых правильный ответ – 1, и алгоритм возвращает 1. Если же объект относится к классу 0, но при этом алгоритм относит его к классу 1 – это «ложное срабатывание» (False Positive). Наоборот, если объект относится к классу 1, а алгоритм относит его к классу 0 – это «ложный пропуск» (False Negative). Если же объект относится к классу 0 и алгоритм его относит к классу 0 – это «верный пропуск» (True Negative).

Данные характеристики отражает матрица ошибок (Confusion Matrix)

	$y = 1$	$y = 0$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = 0$	False Negative (FN)	True Negative (TN)

В случаях, когда разные ошибки имеют разную цену, гораздо лучше измерять другие метрики качества: точность и полноту. Точность (precision) показывает, насколько мы можем доверять классификатору, если он выдает ответ 1. Более формально она равна отношению числа верных срабатываний к количеству объектов, на которых алгоритм выдал ответ 1.

$$precision = \frac{TP}{TP + FP}$$

Вторая метрика называется «полнотой» (recall). Она показывает, как много объектов класса 1 алгоритм находит. Более формально она равна отношению числа верных срабатываний к общему размеру класса 1.

$$recall = \frac{TP}{TP + FN}$$

Итак, точность и полнота характеризуют разные стороны качества классификатора:

- Чем выше точность, тем меньше ложных срабатываний
- Чем выше полнота, тем меньше ложных пропусков

Приоритет в сторону точности или полноты выбирается в зависимости от задачи.

ROC – «receiver operating characteristic». По оси X у этой кривой отложена доля ошибочных положительных классификаций – отношения числа ложных срабатываний к размеру нулевого класса. По оси Y у него отложена доля правильных положительных классификаций – это отношение числа верных срабатываний к размеру первого класса.

Свойства ROC-кривой:

- Левая точка: всегда (0, 0) (все объекты относим к классу 0)
- Правая точка: всегда (1, 1) (все объекты относим к классу 1)
- Если выборка идеально разделима, то кривая пройдет через точку (0, 1)
- Площадь меняется от $\frac{1}{2}$ до 1.

Таким образом, оценивая площадь под ROC-кривой можно оценить качество классификации.

Ход работы

1. Прочитать теоретическую часть по метрикам качества классификации
2. Для двух любых методов классификации из предыдущих работ и своего набора данных посчитать следующие метрики качества:
 - a. Точность классификации (Classification Accuracy)
 - b. Логарифм функции правдоподобия (Logarithmic Loss)
 - c. Область под кривой ошибок (Area Under ROC Curve)
 - d. Матрица неточностей (Confusion Matrix)
 - e. Отчет классификации (Classification Report)
3. Для более точных результатов использовать кросс-валидацию
4. Сравнить применимость используемых классификаторов, основываясь на полученных метриках
5. Отчет оформить в виде блокнота Jupyter Notebook (File → Download as → HTML (.html))